Bergen Journal of Criminal Law and Criminal Justice • Volume 7, Issue 1, 2019, pp. 1-26

A Revisionary Theoretical Framework of Responsibility: A Philosophical Exploration of Incapacity for Responsible Behaviour (*utilregnelighet*)

ATLE OTTESEN SØVIK*

1. Introduction

Both in criminal law and in philosophy, a distinction is made between those who can be held responsible and those who cannot. Those who can legitimately be held responsible are said to have the capacity for responsible behaviour (*tilregnelighet*), while those who cannot legitimately be held responsible lack capacity for responsible behaviour (*utilregnelighet*). But what exactly is capacity and incapacity for responsible behaviour? Is it possible to give a detailed philosophical theory of the content of those concepts which can avoid common objections and integrate common intuitions and practices?

In this article I am interested in *moral* responsibility. This concerns whether someone is blameworthy or praiseworthy for an action, and should be distinguished from *causal* responsibility, which concerns whether someone is the cause of something. We can also distinguish between responsibility as *attributability* and responsibility as *accountability*. Responsibility as attributability is an evaluation of an agent as a moral agent, for example as a saint or a coward. Responsibility as accountability is when a person is treated as someone who should live up to certain expectations and is included in the practice of blame and praise. This is the sort of responsibility I am interested in here.

Norwegian law gives an *extensional* definition of incapacity for responsible behaviour by listing four specific exceptions to capacity for responsible behaviour,

* MF Norwegian School of Theology, Religion and Society.

namely: persons who are under 15, who are psychotic, who have severe impairment of consciousness, and who have severe mental disability (penal code 2005, section 20). However, the law does not give an *intentional* definition of either capacity or incapacity for responsible behaviour, by saying what it means to have or lack this capacity. In legal literature and preparatory works, it is explained as *skyldevne*, i.e., the capacity to be guilty. This capacity is based on certain philosophical premises about the individual's free will and ability to control his/her actions.[1] But how should these philosophical premises be further understood and defended against objections?

In this article, I will present the main features of a theory of responsibility and the conditions that determine who can and cannot be held responsible. It is a philosophical theory, although its point of departure is a problem of law. It is also a consequentialist theory of responsibility, but one which avoids the common objections towards consequentialism. Further details have been developed in two books by the author,[2] but this article focuses on identifying and explaining the main features of different types of incapacity for responsible behaviour (which is not a topic in either of the two books mentioned). I shall argue that we need to understand capacity for responsible behaviour in order to understand the causes of incapacity for responsible behaviour. Further, I will argue that we need to understand the concept of responsibility in order to understand capacity for responsible behaviour. Accordingly, this article will address the following three questions in order:

1. What is responsibility?
2. What is capacity for responsible behaviour?
3. What is incapacity for responsible behaviour?

These are huge questions, and my main focus will be on the third. However, I need to briefly answer the first two questions in order to locate my position, clarify its content, and show how it faces different kinds of objections. I consider it valuable to show the main lines of a theory, even if there is not room in this article to answer all questions or objections.

In the first section of this article, I will start by asking what responsibility is. I will first consider a strict basic desert view that does not allow any consequences to justify what responsibility is, or why we hold others responsible. I reject this view and defend instead a specific consequentialist understanding of responsibility, which I argue is more coherent than the basic desert view. Obviously, there are many possible views that I will not be able to discuss, but the introduction allows me to locate my position and show some of its rationale.

1 Linda Gröning, Tilregnelighet og utilregnelighet: Begreper og regler, appendix to NOU 2014:10, Skyldevne, sakkyndighet og samfunnsvern, published in *Nordisk Tidsskrift for Kriminalvidenskab* 102(2) (2015), p. 409.

2 Søvik, *Free Will, Causality and the Self*, Philosophical Analysis (Berlin: DeGruyter 2016); Søvik, *A Fundamental Theoretical Framework for Science and Philosophy* (forthcoming).

In the second section, I describe in detail the different conditions that are needed for a person to have capacity for responsible behaviour, and also how humans normally develop their capacity for reasoning and making choices. Understanding this is necessary in order to see the different ways in which the development of normal capacity can go wrong and render people incapable of responsible behaviour, which I describe in section three. In section four, some objections are briefly answered.

The article aims to present the main features of a fine-grained theory of responsibility, as distinct from a basic desert view wherein responsibility is just said to be something irreducible. It is a theory that avoids the common challenges from neuroscience which say that nobody is responsible for their actions, and it answers various questions and objections raised throughout the article. Since it is based on a specific and new theory of free will,[3] it is also a new theory of incapacity for responsible behaviour.

2. What is responsibility?

According to philosopher Derk Pereboom, the specific retributivist idea of responsibility as basic desert is the most common and traditional view that is operative in the larger literature, but is rarely explicitly formulated.[4] Manuel Vargas agrees that it is rarely well defined and explained, but refers to Pereboom as the best example of somebody who does it clearly and explicitly.[5] Pereboom argues that responsibility should be defined in terms of basic desert, and offers the following definition of responsibility and basic desert:

> *'For an agent to be morally responsible for an action in this sense is for it to be hers in such a way that she would deserve to be blamed for it if she understood that it was morally wrong, and she would deserve to be praised if she understood that it was morally exemplary. The desert at issue here is basic in the sense that the agent would deserve to be blamed or praised just because she has performed the action, given an understanding of its moral status, and not, for example, merely by virtue of consequentialist or contractualist considerations.'*[6]

The quote explains negatively what it means that the desert is basic by saying what it is not, but it does not positively explain what basic desert is or why the desert is basic. Michael McKenna refers to conversations with Pereboom and explains that, according to Pereboom, there is no justification for this view, as it is simply a basic

3 Søvik (2016).

4 Pereboom, Hard Incompatibilism, in *Four Views on Free Will*, eds. Fischer et al. (Oxford: Blackwell Publishing, 2007), p. 86.

5 Vargas, *Building Better Beings: A Theory of Moral Responsibility* (Oxford: Oxford University Press 2013), p. 251.

6 Pereboom, *Free Will, Agency, and Meaning in Life* (New York: Oxford University Press 2014), p. 2.

relation which cannot be defined in terms of anything more basic.[7]

It should be taken as a warning sign when a theory says that a concept cannot be defined, but is just basic. Often, a theory will say of some concept that is a brute fact or something irreducible, while other theories will say that they can explain it, or that the entity in question does not exist at all, or that the problem of understanding the concept is caused by the wrongful presuppositions of the theory. When something is claimed to be undefinable or irreducible, it is worth looking around to see if anyone else can do better.

In arguments for claims of basicality or irreducibility, appeals to intuitions and thought experiments are not uncommon. Pereboom quotes Immanuel Kant on a thought experiment to defend the basic desert view: Imagine a person who has done something very wrong, and then for some reason turns out to be the last person alive on the planet. This person would still deserve punishment, according to Kant, which shows that responsibility is not just about consequences, but about basic desert.[8]

Maybe many will agree that such a person would deserve punishment, but I do not think that is because there really is such a thing in the world as basic desert. Rather, it is because people assume another understanding of desert than the one that basic desert is meant to express. It could be that they think punishment is good for this person him- or herself, or it feels right because one is not actually able to imagine that everyone else—yourself included—does not exist. We may even think that such a person would deserve punishment because we have a (non-conscious) intuition that good and evil should be distributed evenly in the world.

Concerning basic desert, Vargas refers to Christopher Bennett's social self-governance model of deserving blame. Here the point is that it is crucial to the practice of holding people responsible that people acknowledge that they deserve blame and, due to this, they morally reorient themselves.[9] In other words, it has a good consequence if people believe in basic desert. But all of these reasons are good consequences, and none of them simply say that there is a basic and undefinable relation between a wrong act and punishment in the sense that the act deserves punishment for no other reason than being wrong.

When an idea is not defined, and is not justified by anything other than vague intuitions and thought experiments, we should conclude that we have no good reason to think that something like that exists. This is also the conclusion that Pereboom draws. He defines what the traditional concept of responsibility means in terms of basic desert, then concludes that nobody deserves to be blamed or punished in this sense. The reason for his rejection is that he cannot find any theory of free will to be coherent, and since Pereboom thinks that free will is necessary for basic desert, he

7 McKenna, *Conversation & Responsibility* (New York: Oxford University Press 2011), p. 121.

8 Pereboom (2014), p. 158.

9 Vargas, referring to Christopher Bennett, The Varieties of Retributive Experience, *Philosophical Quarterly* 52(207) (2002).

must reject the latter when he rejects the former.[10]

One could nevertheless argue that people can be responsible and deserve blame or punishment in another sense of the terms 'responsible' and 'deserve'. This raises the following question: if a traditional definition of the concepts of responsibility and desert is found to be problematic, should one then say that responsibility and desert do not exist, or should one find new understandings of the terms?

The many problems concerning the relation between language and the world cannot be discussed here. I follow Quine in thinking that words are interpreted in light of experience, and experiences interpreted in light of words, in a hermeneutic circle aiming for greater holistic coherence.[11] In this process, many pragmatic considerations will influence which words we choose to express which meanings, but a general goal is to limit the revisions for the sake of avoiding misunderstanding. My goal in this article is to make the revisions necessary to have a coherent theory of capacity and incapacity for responsible behaviour, with precise definitions of the terms involved.

Note that I am rejecting basic desert in a narrow sense. People can use terms like consequentialism and retributivism quite differently. While there is much agreement in the extreme ends of the scale of what counts as consequentialism and retributivism, there are many positions in between that could be called either consequentialism or retributivism. For example, consequentialism could be understood narrowly in the sense of a narrow utilitarian justification for punishment, while retributivism could be understood to cover all other theories. Or retributivism could be understood very narrowly as referring only to the basic desert-justification here presented, while consequentialism could be understood very broadly as anything that includes positive consequences of blame. Of course, one can also have different kinds of mixed models in between consequentialism and retributivism. I use the term 'consequentialism' broadly and reject only a narrow understanding of retributivism, and then I use the rest of the article to explain the content of my consequentialist model more precisely. Merely speaking of consequentialism, retributivism or mixed models is too imprecise for my purposes, and one cannot discuss all models in any case, so I have chosen to defend one kind of consequentialism against one kind of retributivism acknowledging that not all models have been discussed.

In the following, I will present the best consequentialist theory of responsibility that I know of, which is developed by Manuel Vargas. He argues that holding others responsible is a general strategy for cultivating morally good agency in a society.[12] We will have to dig more deeply into the concept of responsibility, but a good introduction to this theory can be given by looking at how Vargas responds to common

10 Pereboom (2014).

11 Quine, Two Dogmas of Empiricism, *The Philosophical Review* 60(1) (1951).

12 Vargas (2013).

criticisms of consequentialist theories of responsibility.[13] We start by looking at how Vargas answers four typical objections.

Firstly, we can influence people and animals to make them act better in many ways, but there are many such forms of influence nobody would call responsibility (for example force or bribery), so it cannot be enough to say that responsibility is influencing people to make them behave better—responsibility must be more than that. Vargas answers that holding people responsible is a special form of influence since it is about giving people reasons to act differently, so it applies only to people who can respond to reasons. I will develop Vargas' response by arguing that when people have a normal capacity for reasoning, their behaviour can be influenced by our holding them responsible in such a way that they can implement our responses and see them as reasons for acting in certain ways.

Secondly, it seems that in many individual cases, holding a person responsible will not have the desired effect. He or she need not become a better person by being held responsible. Vargas agrees, but points out that holding others responsible is a general strategy for cultivating morally good agency in a society, and that it obviously works at that level, even if it does not work in many individual cases.

Thirdly, Vargas' theory seems not to distinguish between holding people responsible and holding people appropriately responsible. Here the objection is that if holding people responsible is just about influencing their behaviour, it seems we could imprison innocent people as scapegoats in order to achieve the desired result (that people behave better). Since it seems inappropriate to punish an innocent person, the objection is that a theory of responsibility must include more than just influence to explain when it is appropriate to hold others responsible. Vargas answers again that a consequentialist theory of responsibility must be seen as a general way of cultivating good agency in a society, and as a general strategy, it will not work to imprison innocent people.

In addition to Vargas' response, there are other ethical norms that are reasons not to imprison innocent people: that it is not just, for example, or that it is not the best way to actualise the best world. I write more on this in section four, objection one. The point I am making here is that a theory of what responsibility is will not answer all questions related to responsibility, since the same theory of responsibility can be combined with different ethical theories. In addition to a theory of responsibility, we need further arguments to say something about hard versus soft punishments, and why blaming innocent people is not the best way to actualise the best world. As such, factors other than the theory of responsibility itself come into play when it comes to determining how, or in what way, and in what degree people should be held responsible.

Fourthly, it seems that we often blame people without being interested in influencing them. We may even blame dead people, even though influencing them is im-

13 Idem., pp. 187-95.

possible. Vargas responds that others can learn moral lessons when we blame dead people, even if the dead cannot. Vargas acknowledges that, in many cases, people will not have the intention of cultivating agency when they blame people, but that nevertheless the whole practice of holding each other responsible has this effect.

To sum up so far, Vargas argues that we hold others responsible as a general strategy for cultivating morally responsible behaviour. But does not holding someone responsible presuppose them being responsible? We have seen that in the philosophy of responsibility there is a distinction between basic desert views and consequentialist views. These two views have a very different understanding of the relation between a person having capacity for responsible behaviour and a person being held responsible. The basic desert view sees the capacity for responsible behaviour as primary, and if this capacity is present and a person does something wrong, then it follows that the person deserves blame or punishment, regardless of what consequences blaming the person has. The consequentialist view sees holding others responsible as primary. Holding others responsible by praising, blaming or punishing them is what creates in them the capacity for responsible behaviour. It gives them an understanding of the world and appropriate feelings about what is good and bad, in light of which they can guide their behaviour.

Briefly put, the first view says that only if you have capacity for responsible behaviour can you be held responsible, while the other view says that you can achieve capacity for responsible behaviour by being held responsible. But even to shape responsible persons by holding them responsible does require a normal capability for thinking and feeling. Even if one thinks that holding persons responsible is primary in understanding what responsibility is about, one could still say that it presupposes a basic capacity for responsible behaviour. This capacity then means that is possible for the person to be influenced by being held responsible through a normal deliberation process. In the following, I will try to describe in more detail how this happens.[14]

What usually happens when we hold others responsible is that we compare a person's action with a moral standard concerning what a person should have done in such a situation. This means that you can also be held responsible for something that you have not done, but should have. For example, if a child is drowning while you are nearby and you could easily save it, people think that the morally right thing to do in such a situation is to jump into the water and try to save the child. If you do not, people will hold you responsible and blame you according to this standard, even if you did not cause the child to drown. If you fail to do what people think you should have done, or do what they think you should not have done, it is considered blameworthy. Conversely, acting in accordance with the moral standard in these situations is considered praiseworthy.[15] For this reason, I think that one can be morally responsible for something one is not causally responsible for, contrary to what, for example,

14 Based on Søvik (2016).

15 Bok, *Freedom and Responsibility* (Princeton, NJ: Princeton University Press 1998).

Michael McKenna holds.[16]

People are usually understood to have a normal capability for formal reasoning (basic ability for induction and deduction, even if it is also common to make errors) and a normal emotional life, which means that there are many standard scenarios that will make people happy, sad, angry, feel pain, etc. When we hold people responsible, we assume that people understand that doing this or that standard action will hurt people, make them angry, make them happy, and so on. Because people usually have normal capacities for reasoning and experiencing emotions, other people expect that they will have a normal understanding of what is true and what is right to do in various situations. We expect that they learn certain norms and that these norms are alternatives they consider when they make choices. As people grow older and become adults, we expect that they have had enough time to think and to make their own experiences such that they can understand essentially what is true and right. We also expect them to act in accordance with that understanding.

What does it mean to be responsible (in the sense of having capacity for responsible behaviour) as opposed to being held responsible? In the theory offered in this article, a person is responsible for what he or she does or does not do in a situation if it is type—not token—physically possible for that person to act in a morally different way, in a way which can be influenced by being held responsible.

I will start to unpack the content of the previous sentence by introducing the type-token distinction: a type is defined by its properties regardless of space and time, while a token is also defined by its location within space and time. If I say 'chair, chair, chair', there is one type and three tokens. If two objects or events are completely similar to each other, they are one type and two tokens. They are one type because they have the same structure, but they are two tokens because they are located at different places in space and/or time. This distinction is important for discussions about the possibility to act otherwise in a situation, since in a specific situation there may just be one action that is token physically possible for a person to do given the causes that led up to the situation. Regardless of whether the world is determined or not, many neuroscientists will say that processes in the brain controlling our actions happen in a causal chain as if they were determined. Nevertheless, it makes sense to hold persons responsible in such cases because the action of holding someone responsible can itself influence what is token physically possible to do in a specific situation. Why is that important?

The relevance of this point is that many (typically physicists or neuroscientists) will argue that no humans are responsible for their actions since all our actions are results of causal chains in the brain where it never makes sense to speak of an agent

16 McKenna (2011), p. 7.

considering alternatives and controlling the outcome.[17] There is just one alternative action that it is token physically possible for a person to do in a specific situation because of the causal chain that led to this action. But one may accept that and reply that holding others responsible is a part of the causal chain that led up to the action and influences exactly which alternative is the token physically possible one. Jack seeing the telling look from his father may cause another alternative action to become the one that Jack's brain executes.

We need the discussion of type and token possibilities in addition to the discussion of determinism, because we can distinguish between two levels: the brain, and the world as a whole. If the world is determined, there is no point in trying to change the future. But if the world is not determined, the brain may still work as if it was determined, meaning that only one action was token possible to perform at a certain time. However, if the world is not determined, we can influence and change what the only token possibility is for a person at one specific time (something we could not do if the world was determined).

If the world is determined, there is only one token physically possible chain of events that can happen from the Big Bang until the end of time, and then it does not make sense to say that holding others responsible is a meaningful way to change what is token physically possible in a specific situation. Do we have reason to believe that there is indeterminism at the macro level of human interaction? While a full discussion of determinism and indeterminism is not possible here, I will offer some arguments to support that we should think that the world is not determined. The most common place to go for support is quantum mechanics. Quantum mechanics can be given indeterministic interpretations (like Copenhagen and GRW) and deterministic interpretations (like deBroglie-Bohm and Everett). Nevertheless, all interpretations will agree that the guiding laws are merely probabilistic, saying only that something will occur with a certain probability.[18] This still leaves open whether there is a determinism at a deeper level, and whether indeterminism at the micro level of elementary particles can be scaled up to the macro level of human interaction.

If there is indeterminism at the micro level of quantum mechanics, there may nevertheless be determinism at the macro level, because the events at the micro level cancel out at the macro level. It may be undetermined whether a single particle goes

17 See for example Wegner, *The Illusion of Conscious Will* (Cambridge, MA: MIT Press 2002); Pereboom (2014); Singer, Verschaltungen legen uns fest: Wir sollten aufhören, von Freiheit zu sprechen, in *Hirnforschung und Willensfreihet. Zur Deutung der neuesten Experimente*, ed. Geyer (Frankfurt am Main: Suhrkamp 2004); Gazzaniga, *The Ethical Brain* (New York: Dana Press 2005); Haggard, Decision Time for Free Will, *Neuron* 69(3) (2011); Harris, *Free Will*, (New York: Free Press 2012); Honderich, *A Theory of Determinism: The Mind, Neuroscience, and Life-Hopes* (New York: Oxford University Press 1988); Hawking and Mlodinow, *The Grand Design* (New York: Bantam Books 2010); Einstein, quoted in Frankenberry, *The Faith of Scientists in Their Own Words* (Princeton: Princeton University Press 2008), p. 145.

18 Ney and Albert, *The Wave Function: Essays on the Metaphysics of Quantum Mechanics* (Oxford: Oxford University Press 2013).

here or there, but determined that 50% will go here and 50% will go there, so that the macro result is the same in any case. If it is undetermined where the particle will go, we could set up contrasts and ask 'why did the particle go here as opposed to there?', and the answer will be that there is no cause, but rather it is a causeless undetermined event. However, it could also be that undetermined micro events can scale up to the macro level.

James Ladyman offers the example of a scientist who decides to take lunch after so and so many clicks on her Geiger counter.[19] Geiger counters measure events that, according to some interpretation, are indeterministic. We could expand the example and say that a scientist may decide to invite her male colleague to lunch if she gets a click on her Geiger counter before 12. This decision may cause them to have lunch, fall in love and get married—or not. The world may then be very different in the future depending on undetermined events. We do not know whether quantum mechanics should be interpreted deterministically or non-deterministically. But note that also in Newtonian physics indeterminism at a macro level can occur, for example if three identical particles with the same speed collide.[20]

Here is an additional argument I suggest in favor of indeterminism: evolution makes more sense if it is genuinely open what the content of the future will be than if it is determined at the micro level. If many scenarios are possible, we can easily understand why the one that was actualised was the one where the best fit for survival produced many children. If the one scenario that actually happened was determined solely by laws governing particles at the micro level, it seems that this scenario could just as well have been a scenario where what happens at the macro level is very chaotic and unsystematic. The selection effect makes more sense as a selection between genuinely possible futures than if only one future was possible anyhow.

Regardless of the discussion on determinism, the lawmakers have nothing to lose and everything to win by presupposing that the world is not determined. For if law presupposes that the world is determined, blame and punishment do not make sense as attempts to change future behaviour, since the outcome is determined anyway. If the lawmakers presuppose that it is not determined, then either they are right, and the law makes sense, or they are wrong and the world is determined, in which case the lawmakers just did what they were determined to do. Indeed, the law presupposes that the future is open, since it tries to make people behave in one way rather than another. This concludes my discussion of why we have reasons to believe in indeterminism.

I said above that there exist no entities in the world to which the concepts of basic

19 Ladyman et al., *Every Thing Must Go: Metaphysics Naturalized* (Oxford: Oxford University Press 2007), p. 264.

20 Earman, *A Primer on Determinism*, University of Western Ontario Series in Philosophy of Science (Boston, MA: D. Reidel Pub. Co. 1986), pp. 30-32. Earman also gives other examples from Newtonian and relativity physics. Important examples are briefly summarised in Sklar, *Philosophy of Physics*, Dimensions of Philosophy Series (Boulder: Westview Press 1992), p. 203.

responsibility or basic desert refer. This is contrary to those who think of these concepts as referring to irreducible entities, as presented in the beginning of this section. In my view, what exists are people who influence each other by comparing actions with a moral standard for what they think should have been done in that situation, and then praising, blaming or punishing people accordingly, which makes people consider such reactions when deliberating. This practice only has the effect of cultivating moral agency in cases where people have the capacity for considering such reactions when deliberating. Thus, we may define capacity for responsibility as capacity for considering such reactions when deliberating.

The standard with which we compare the actions of people is a general standard that presupposes normally developed people with normally functioning minds. But there may be many different developments or functions of the mind that are not normal, and where attempts at moral influence in light of a standard does not have the desired effect. In the next section, I shall consider in detail the normal development of a person and a normal deliberation process, which are what constitute capacity for responsible behaviour. This will then be used in the last section to explain the different cases of lack of capacity for responsible behaviour.

I end this section on responsibility with some final distinctions concerning the meaning of 'being responsible'. When it comes to being responsible, we should distinguish between being a responsible person in general and being responsible for x in a particular situation. Being a responsible person in general means that you have capacity for responsible behaviour and can take praise and blame into consideration in a normal process of deliberation in any situation where it is possible to deliberate. Being responsible for x in a particular situation means that there is something (x) that a person has done or not done in this particular situation which could have been influenced by praise or blame through a normal process of deliberation.

Being blameworthy or praiseworthy for x in a particular situation means that there is something (x) that a person has done or not done in this situation which, according to a moral standard, is blameworthy or praiseworthy. Responsibility in itself is nothing more than a concept referring to people being responsible for something in a particular situation, which again is reducible to a certain situation involving a person with capacity for responsible behaviour, and 'capacity for responsible behaviour' means the capacity for being influenced by praise or blame through a normal process of deliberation. All these concepts have content which refers to normal processes between humans. There is nothing mysterious, unknown or metaphysical, in the sense of being non-empirical or irreducible or undefinable, about them. It is now time to look, in more detail, at how to understand this normal process of deliberation which constitutes capacity for responsible behaviour.

3. What Characterises a Person with Capacity for Responsible Behaviour?

In this section, I want to describe in detail what happens when a person makes a choice, and how a person normally develops what we call free will. Free will can be defined in many ways, but most philosophers will have an understanding of free will and responsibility where free will (as they understand free will) is presupposed as necessary for responsibility. This includes myself, so I will offer an account of what kind of free will is necessary in order to have capacity for responsible behaviour. It will be a detailed theory which relates traditional concepts to scientific concepts with empirical content. In section 3, I will show how it explains incapacity for responsible behaviour.

In this section, I will start by describing how to think of a choice as a causal process in the mind. I move on to show different types of choices, characterised by different degrees of involvement by the person making a choice. I then move on to describe how a person develops a self, which can be independent (self-governing) to different degrees. Two important insights that will follow from this are that freedom and responsibility come in degrees, and that different factors influence the degree of responsibility.

Even if I reject compatibilism, my theory is similar to many compatibilist theories in its understanding of responsibility and alternative possibilities, and in its focus on how the agent is the source of her action in an event-causal way (which means that causes in the mind are like other causes in nature). It is especially close to the compatibilist theories called source compatibilist theories or quality of will theories. The crux of these theories is that an agent in some sense is the source of the choice, or that the action expresses their attitudes, and that this is what it takes to have free will.[21] A classic theory here belongs to Harry Frankfurt, who argues that we have several desires whose object is an action or a state, which he calls first-order desires. But, we also have desires whose object is a first-order desire, and these he calls second-order desires. The second-order desires are internal responses to the first-order desires, which one may like or dislike. According to Frankfurt, we are free when our second-order desires approve our first-order desires, because only then do we have the will that we want.[22] Closest to my theory is probably the deep self theory of Chandra Sripada, in which we are morally responsible for the actions that express our deep self, and our deep self is our set of cares.[23] I think differently from Sripada when it comes to what the self is and how it causes actions. Most importantly, the difference between my theory on the one hand, and the deep self theory and other

[21] See for example Smith, Control, Responsibility, and Moral Assessment, *Philosophical Studies* 138(3) (2008), or Scanlon, *Moral Dimensions: Permissibility, Meaning, Blame* (Cambridge, MA: Belknap Press of Harvard University Press 2008).

[22] Frankfurt, Freedom of the Will and the Concept of a Person, *Journal of Philosophy* 68(1) (1971).

[23] Sripada, Self-Expression: A Deep Self Theory of Moral Responsibility, *Philosophical Studies* 173(5) (2016).

compatibilist theories on the other, lies in how I use a contrastive theory of causation to show how the self can be an ultimate cause of an agent's actions.[24] Here I differ from all compatibilist theories in arguing that we have free will in a stronger sense than compatibilists admit, which is incompatible with determinism.

How can a choice be understood as a causal process in the mind? Here is one example: A person sees a chair, and a woman approaching. The visual impression of the chair and the woman activates in him two desires: a desire to sit and a desire to get to know her. The same visual impression also activates fact memories like 1) the fact that he can offer the chair in order to get to know the woman, and 2) that it would be considered rude to sit down on the chair in front of the woman. These fact memories further activate autobiographical memories of being rude previously and the bad feelings connected with that, and of being polite to women before and the good feelings connected with that. The good feelings remembered are connected with the desire to offer the chair, and the bad feelings with the desire to sit. The desire to offer the chair becomes the strongest and activates the motor neurons that make the man offer the chair. In this process, every event was causal, because it was all about neural patterns firing and thereby activating other neural patterns with which they were connected. The process could also be described as a man acting for a reason. He offered the chair, and the reason was that he wanted to get to know the woman since that was his strongest desire.

The example shows that acting for reasons does not exclude acting for causes. A causal understanding of the mind is the orthodox view in philosophy of mind.[25] All non-causal theories of the mind struggle with a challenge from Donald Davidson, namely to specify what it means to decide for a reason if the reason is not understood as the cause of the choice. We have different desires, thoughts and motives in mind when we deliberate on how to act, but how can any of these lead to action if not by causing the action? In this article, I defend a theory of choices where the strongest desire leads to action, as I have discussed elsewhere.[26]

Stated differently, what we usually describe as deliberating over different alternatives can more coherently be understood as different alternatives being activated and becoming conscious in a causal process. What feels like making a choice is, in fact, a causal process where an input from the body or the world external to the body activates beliefs from memory about alternatives for action. These beliefs activate desires, which interact with the beliefs, causing one of the desires to reach a strength above a certain threshold, thereby activating the body to move. This is a finely-grained and coherent description with much empirical support, which will be

24 For more details on this contrastive theory of causation, see Søvik (2016), chapter two.

25 Mele, *Irrationality: An Essay on Akrasia, Self-Deception, and Self-Control* (New York: Oxford University Press 1987), p. 31.

26 Søvik (2016). Thanks to Jørn Jacobsen for showing me that Francis Hagerup has a very similar description to mine of the deliberation process as a causal process, see Hagerup, *Strafferettens Almindelige Del*, vol. 3 (Oslo: Norsk bok-duplisering (Johs. Minsaas) 1930), p. 4.

developed further below. In comparison, the alternative theory that an agent with free will makes a sovereign choice between alternatives is very problematic. What is meant by 'agent', 'will', and 'freedom' here, and how does such a choice come about? By what criteria is the choice made; where did these criteria come from; and why does this scenario make a person free?

When a person (soon to be defined further) makes a choice, the person may be involved in the choice to different degrees. This may seem overly detailed, but it is important in order to understand moral excuses in the next section. I start by mentioning some actions that are not caused by desires at all, such as breaking a glass by accident or just having a reflex movement. Such actions should probably not be called actions at all, since the concept of action should include a desire. The next processes described do involve choices, and constitute different kinds of deliberation processes.

At the first level of action, we find actions caused by desires. The desires can be conscious or non-conscious, innate or acquired. Sometimes desires lead directly to action without the occurrence of any additional thoughts or feelings between the desire and the action happening. For example, a man might see someone being rude to his girlfriend, desire to hit that person, and then immediately do so.

A new level of personal involvement is when the autobiographical self of a person is activated between the desire occurring and the action happening. The autobiographical self will soon be defined, but for now, we can think of it as a repository of important memories in the brain. Sometimes the autobiographical self is activated and changes the initial desire. Consider the example of the man who desires to sit but also sees a woman approaching the same chair. Autobiographical memories of not having offered a seat to someone previously and then receiving negative feedback are activated, along with others of having offered a seat and the positive consequences of this. Then the initial desire to sit weakens, and the desire to offer the seat to the other person strengthens. Initial desires may also change because of new thoughts and feelings, as the brain is capable of making new connections between neurons.

The autobiographical self can be more or less involved in deliberations done between an initial desire and an action, depending on both the number and the emotional strength of the memories considered before acting. The autobiographical self may also be more or less independent. An autobiographical self becomes increasingly independent throughout life as it changes initial desires through a process of thinking and feeling about alternatives. Even if the autobiographical self does not decide what to think or feel, such a change in initial desires is nevertheless caused from within the mind.

These were different levels of involvement in a choice, but they referred to an autobiographical self (which again can become more and more independent). This will now be further explained. I use the terms 'person' and 'agent' interchangeably for a living human body with a mind and a core self. The concept of a core self comes from the neuroscientist Antonio Damasio's understanding of the self. Damasio ar-

gues that the self was built in three different stages.[27] The first stage is what Damasio calls the proto-self, which is a neural pattern representing the whole organism. This proto-self produces a primordial feeling, which is the feeling of my own body existing, but without any further connection to the world.

In addition to the proto-self, the brain creates neural patterns representing objects and events in the world, but it also creates neural patterns representing the relationship between the organism and the outside world. From moment to moment there is a series of neural patterns representing how the organism changes in relation to the outside world. This creates changes in the primordial feeling, which are consciously felt as an experience of changes in the world. The representations of change create pulses of core consciousness that together constitute the core self, which is the second stage.

Finally, these conscious experiences can be held together in extended consciousness to create the autobiographical self, which is a neural pattern representing the life story of a person, created by memories and continuously reconstructed. Memories of past experiences constantly return to the mind and influence what happens in the mind when presented with new choices. A person has experiences which lead to feelings and thoughts that are stored in the memory. The more memories of thoughts and feelings a person has connected to her experiences, the more these memories will influence what the person desires later, since desires depend (among other things) on what we feel about different alternatives. The autobiographical self is a collection of memories of thoughts and feelings which influences the desires and choices of every person. It becomes an increasingly significant influence on most people's choices during their lives, since an increasingly large collection of thoughts and experiences can be remembered, and these influence future choices.[28]

Even if nothing can be the cause of itself, an autobiographical self can, over time, be the cause of its own content. When we start our lives, we are not free. At the beginning of life, children follow their initial desires and are told by their caregivers who they are, and what is right or wrong, and good or bad. But most children are born with a capacity for reasoning, feeling, thinking new thoughts, and making new connections in their minds, which gives them a general ability to find out what is true and good and right. When they make choices and have experiences, these are added to their autobiographical selves.

Later, they experience new processes in which new thoughts and feelings change

27 Damasio, *Self Comes to Mind: Constructing the Conscious Brain* (New York: Pantheon Books 2010), chapters 8 and 9.

28 Let us say that future research in neuroscience rejects Damasio's distinction between the core self and the autobiographical self. There will nevertheless remain something structurally similar, where something corresponds to our conscious experience of here and now, while something corresponds to the important memories that shape how we experience ourselves as persons and what we desire. These parts of a better theory of the self in the future must then replace what I here call the core self and the autobiographical self.

the initial desires and cause action. Again, the autobiographical self is changed from within the mind of the person. In later choices, the old choices and experiences from the autobiographical self can change the initial desires, and again the experience is stored in the autobiographical self. So, sometimes a mind-internal process happening independent of the autobiographical self can change the autobiographical self, for example, if a new experience gives rise to a new thought or a new feeling. But most often when the autobiographical self changes, the process will involve the autobiographical self, such that it causes choices that cause changes to the autobiographical self.

This means that as long as the world is not determined, it will often be right to select the autobiographical self as the cause of a change in initial desires *and as the cause of new changes in the autobiographical self*. This is illustrated in Figures A and B.

Figure A. Input comes through the senses, and thoughts and feelings influence what is desired most strongly. The person acts and the action is stored as a remembered experience in the autobiographical self.

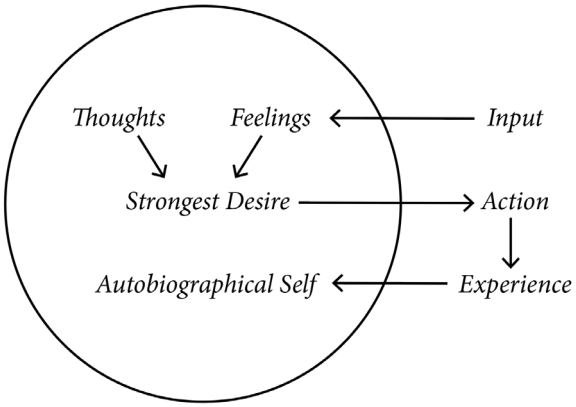
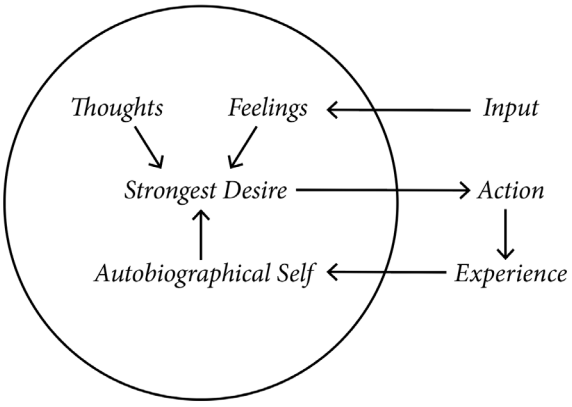


Figure B. Input comes through the senses, and thoughts and feelings influence the desire. The autobiographical self also influences which desire becomes the strongest. The person acts and the action is stored as a remembered experience in the autobiographical self. In this way, the autobiographical self can cause changes to itself over time.



This model explains how holding others responsible influences their choices. Being held responsible is an experience which is stored and influences later choices. This may seem to lead to a regress problem, since presumably there must be some causes external to the self determining how the autobiographical self becomes what it is. The solution to this problem requires a discussion of causality, which there is not room for here, but I discuss causality and this regress charge at length elsewhere and conclude that even if we have a non-chosen starting point, in an undetermined world it will often be right to select the autobiographical self as the most important cause of an action.[29] With this theory in hand, it is now possible to translate traditional concepts and give a more detailed content to what is meant by influencing a person, a person taking blame and punishment into consideration, or a person changing a desire, and how all this relates to the concept of control. This will be further unpacked now.

A person has capacity for responsibility if he or she can take praise and blame into consideration in a normal process of deliberation. There are many different normal processes of deliberation, but a basic part of them is that they activate alternatives for actions in our mind to choose from, where one alternative has the strongest desire connected to it and leads to action. This is a choice, and the choice is freer and more responsible to the degree that it is caused by a self-formed self.

Free will and responsibility are not either-or issues. Instead, each of them exists on a continuum, where the degree of freedom has to do with the involvement of the autobiographical self—how strongly it is involved in the deliberation process, and how independent it is by having been involved in earlier deliberation processes. The autobiographical self being the cause of a person's actions is the content of the term 'self-control.' What it means to control one's actions is to cause one's actions (as argued by Alfred Mele[30]), for how can you have control over something to which you are not causally linked? In order to have an effect on something, one must be causally linked to it. What it means to control one's action is simply that the autobiographical self causes it. The degree of control that we have over our actions is the degree to which our actions are influenced by an independent autobiographical self.

The way that we influence a person, by holding him or her responsible, is by praising and blaming or punishing in different ways. The way that this influences the person is that (expected) acts of praise and blame are events the person experiences and stores in their memory, which is then activated and influences future choices. The activation of this memory and its influence on choices is what it means to say the person takes the praise or blame into consideration. The influence on the choice is that a memory can cause a desire's strength to change.

What I have tried to show here is that it is possible to translate a vague theoretical framework (which could be exemplified by normal daily speech about persons

29 Søvik (2016), chapter two.

30 Mele, *Autonomous Agents: From Self-Control to Autonomy* (New York: Oxford University Press 1995), p. 10.

having free will and control over which alternative they choose) into a much more detailed theoretical framework, where all the concepts have clear empirical content. Next, I shall apply this theoretical framework to analyse and discuss traditional cases where we think people lack capacity for responsible behaviour, or have a diminished responsibility.

4. What is incapacity for responsible behaviour?

Capacity for responsible behaviour is the ability to take praise and blame into consideration in a normal process of deliberation, which was explained earlier. When we think that this ability is sufficiently developed, we say that persons have capacity for responsible behaviour. However, many parts of this capacity come by degrees, and there are many ways in which the process of developing and maintaining such a capacity can go wrong. There may, therefore, be many kinds of mitigating excuses for blameworthy actions and, at a low enough level, we may say that people do not have capacity for responsible behaviour at all. In the following, I will use the presentation in section 2 of how a person develops the capacity for responsible behaviour as an outline to explain many common examples of reduced responsibility.

In section 2, I described how people can be involved in their choices to different degrees. I started with 'actions' which were not caused by desires at all, such as reflex movements or breaking a glass by accident. Normally, we do not hold people responsible for such actions or even call them actions at all. This makes sense, since in such cases there is no deliberation process that can be influenced by holding them responsible.

There may be cases where we do hold people responsible for accidents not caused by desires, but in those cases what we blame them for is not preventing the accident in the first place. For example, if a person, carelessly listening to music, walked through a shop with very expensive glasses and broke a glass, we may blame them for not walking more carefully in order to prevent accidentally breaking glasses. But in clear cases of accidents, we usually do not blame people.

The next level of involvement I described were actions caused by desires, but where the autobiographical self was not involved, such as if a person hit another person before having time to think about it. The law more strictly punishes premeditated violence than violence that happens spontaneously. This makes sense, since in planned actions, the autobiographical self has had plenty of time to prevent the person from doing the act. If it has not prevented it, then the autobiographical self should be changed, and punishment can have that effect.

In situations where a desire causes action without the autobiographical self being involved, we still blame people. But why does that make sense if the autobiographical self has not been involved to prevent the action from happening? Here it is rel-

evant to distinguish between desires that are changeable to different degrees. Some desires are impossible to change, such as the desire for water or the desire for sleep, so we will never blame someone for having these desires. Some desires are possible to change, such as the desire to hit someone because they stepped on your toe, or because they have black skin, etc. We want to influence your autobiographical self to change those desires so that they do not occur at all, or at least so that they are immediately prevented by the autobiographical self from leading to violent action. In other words, we want you to form your character into becoming a person who does not suddenly hit someone unless there is a very good reason. However, some of these desires are particularly difficult to change. For example, in cases of addiction or using violence if you suddenly catch someone abusing your child, the difficulty of changing certain desires may be mitigating.

The clearest cases where we hold others responsible are when they make choices where an independent autobiographical self has been developed and has been involved in the choice. Developing an independent self takes time, but the Norwegian lawmakers have decided that when people are 15 years of age they can be held responsible, implying that by then they should have had enough experience and time of reflection to learn and internalise the basics of what is true and false, good and bad, right and wrong. People have different starting points in life and different conditions for learning, but after 15 years of experience it is common to think that all should have learnt the basics well enough that we consider it appropriate to use punishment to influence their behaviour. Whilst some countries have chosen ages other than 15, and no exact age can be set as correct for all because of the differences between individuals, we can at least say that the appropriate age of responsibility is not very different from 15, at least if everything else is normal.

However, there may be special circumstances where 15 years is not enough to develop an independent and normally functioning autobiographical self. A traumatic childhood or growing up in a brainwashing sect may have disturbed this development, and these backgrounds are, at any rate, very different from the normal conditions that we assume when comparing the actions of a person with a standard for what that person should have done. This may then be mitigating in cases where they have done something wrong.

Being less than 15 years of age is one of the four cases where the Norwegian penal code says that persons lack criminal capacity. This case is quite different from the three others. It has to do with developing an autobiographical self, which clearly happens gradually. The three other cases (psychosis, severe impairment of consciousness, and severe mental disability) are quite different. Persons with these conditions lack some basic features of a correct understanding of the world, and/or their formal capacity for reasoning (drawing correct conclusions from premises) does not work properly. All people have some errors in their understanding of the world, and all people can make bad inferences, but this comes in degrees. In extreme cases, such as those listed above, this understanding can become so poor that blame and pun-

ishment do not—considering the condition of such persons in general—have the intended effect.

If the court decides to punish a person, but the person thinks that the judge in an evil alien in disguise, or is unable to understand the connection between punishment and actions, then blame and punishment do not have the intended effect. I am not an expert on the diagnoses of psychosis, severe impairment of consciousness, or severe mental disability, but in this article I have tried to offer a theoretical framework and conceptual apparatus that can be used as tools by those who are experts on these diagnoses, but who are not experts on the conditions for having capacity for responsible behaviour.

This conceptual apparatus could help forensic psychiatrists to explain, for example, that a person has a cognitive disability which makes him or her unable to draw an inference from 'something causing unnecessary pain for others' to 'that it should not be done', or to draw an inference from 'I will be punished for this' to 'I should not do this'. In other words, states or conditions wherein the normal process of deliberation does not work. Holding such persons responsible will not have the intended effect and will be pointless. Another example could be to explain that a person has a psychosis which makes him or her unable to distinguish between which states of affairs are part of the real world and which are not, and is thus unable to draw inferences from real states of affairs only, which again means that the normal process of deliberation does not work at this point. Here too, holding such persons responsible will not have the intended effect and will be pointless.

Even if I do not know the details in the psychology, it seems that this kind of approach to discussing responsibility in court would be much more fruitful than, for example, the discussions of diagnoses that we saw in the Breivik trial. Just establishing a diagnosis will often not be enough to say whether the person in question had the capacity for responsible behaviour.

The malfunction of the capacity for a normal process of deliberation that we see in psychosis or cognitive disability is different from the legal presumption of insufficient capacity in people under 15 years, since in their case they have not yet been able to build a sufficiently robust capacity for normal deliberation in moral issues. This is because it takes time to learn what is right and wrong, and to form a character where reasoning that something is wrong also causes you not to act wrongly. We give people time to establish this capacity on their own, but if they have not been able to do so by the age of 15, the threat of punishment is there to help them in the right direction.

Holding someone responsible is an act which has the purpose of influencing a normal process of deliberation and thus presupposes a capacity for that process. If this capacity is lacking in a person, then holding this person responsible is inappropriate, since the conditions are not met. This allows us to highlight the difference between someone not having capacity for responsible behaviour and someone not being punished. Someone may have a capacity for responsibility, yet we do not pun-

ish them for some reason or another (e.g., they had a good excuse). The cases I am interested in here are those in which a person is not punished because the person in question does not have the capacity to take the punishment into consideration in a normal process of deliberation. Whether they have this capacity or not is then the relevant question.

The basic idea, that punishment is only appropriate when it can influence a normal process of deliberation, is to be understood at a general level and not as something considered for each individual. This is because the actual effect of punishment on one specific individual is influenced by so many complex factors that we can never know for sure the effect in one specific case. But we do know in general that (threat of) punishment has a corrective effect on behaviour—the threat of fines, for example, causes more people pay their tickets at the subway.

This basic idea explains why we would not blame or punish someone that did something wrong under the effects of hypnosis, or if we learned that their brain was controlled by an evil scientist. Then we would blame instead the hypnotizer or the evil scientist, since the purpose of blame is improved behaviour and the hypnotizer or the evil scientist is the cause of the bad behaviour. The same applies to cases of people acting wrongly under duress, unless we think the person being forced to act should have resisted the force.[31]

It is a challenge that all the parts of a normal development of an autobiographical self and all the parts of a normal deliberation process vary in degrees. Consciousness does not vary in degrees, but on the model described here, it makes sense to blame people for non-conscious acts as well, if we think that their autobiographical self should have prevented such actions from occurring. For example, I can blame you for not putting the toilet seat cover down even if you did not consciously leave it open. The autobiographical self can cause actions consciously and non-consciously, and that is not an important difference, although a conscious choice is usually a choice where the autobiographical self is more involved. What matters is the involvement of the autobiographical self, and not whether it is conscious, since we have no good reason to think that consciousness has a causal effect in any case. This insight is important in understanding why the classical objections from neuroscience pose no threat to the concepts of free will and responsibility. I return to this below.

I have already mentioned that desires are changeable to different degrees. The capacity persons have for formal reasoning also comes by degrees and is closely linked to IQ. This is reflected by the fact that the state defines the level of IQ needed in order to have criminal capacity. Another tricky question is what to think of degrees of knowledge that people should have. Lack of knowledge is mitigating in various cases

31 By 'force' I mean external force. A feeling of internal force making you do something you do not want to do is a sign of a malfunctioning mind, and then it must be considered whether it is. Since all mental states are caused, there is no meaningful distinction to be made between internally forced and unforced mental states.

unless we think that the person should have known the relevant matters and can be blamed for the lack of knowledge. For example, I will not blame a person for not trying to save another person's life by use of an advanced technique which that person has not heard of and few others know. But I might blame a person for not trying basic CPR to save a person and, even if the person does not know CPR, I might blame him or her for that, since I think that this is important knowledge that people can be expected to have heard about, should want to learn, and should actually be expected to have learned.

A person who has a very strange opinion of the world may be to blame for it if we think that this person has only searched for information in strange places and should have been more self-critical. Even if there are many complex cases of what people know or should have known, it is very helpful in the evaluation of these cases to see that we base our evaluation on a comparison with what we think people normally should have done in and before a situation. We can then specify exactly what we think people should have done and consider whether they had the capacity to do this.

In addition to knowledge and desires, our actions are also influenced by our emotions (since they influence desire strength),[32] and we hold people responsible by comparing their actions with a standard that presupposes a normal emotional development. What then to think of people who lack empathy for biological or genetic reasons, the so-called psychopaths? Psychopaths seem able to detect moral reasons cognitively without recognising their moral force emotionally. Some philosophers argue that detecting moral reasons alone is sufficient for responsibility,[33] while others argue that recognizing their force is also required.[34] According to the theory presented in this article, it will again depend on the degree to which they can change their behaviour through normal deliberation as a result of being held responsible. If (the threat of) punishment can influence their reasoning and prevent them from doing

[32] For details on this, see Søvik (2016).Very briefly put, Alfred Mele defines a desire for A as an A-focused attitude which constitutes motivation for A (Mele, *Motivation and Agency* (New York: Oxford University Press 2003), p. 170.). It includes both thoughts that something is good and a feeling that makes the person want the desired state of affairs to happen or be true. Antonio Damasio distinguishes between emotions and feelings. An *emotion* is a series of events happening in the body, where various chemical molecules are sent out in the blood and different signals sent through the nerves. A *feeling,* on the other hand, is a neural pattern in the brain representing the body being in that state of emotion. It is a neural pattern representing the body being, for example, in a state of fear or happiness or anger, which we can feel consciously as being afraid, happy or angry (Damasio, pp. 109-14). When we experience something as feeling good or bad, this will influence how strongly we desire the same to happen in the future.

[33] Scanlon (2008); Talbert, Moral Competence, Moral Blame, and Protest, *Journal of Ethics* 16(1) (2012).

[34] Levy, *Consciousness and Moral Responsibility* (Oxford: Oxford University Press 2014); Shoemaker, *Responsibility from the Margins* (Oxford, United Kingdom: Oxford University Press 2015); Watson, The Trouble with Psychopaths, in *Reasons and Recognition: Essays on the Philosophy of T.M. Scanlon*, eds. Wallace, Kumar, and Freeman (Oxford: Oxford University Press 2011).

wrong, then it seems that their capacity for deliberation is normal enough that they should be held responsible for their actions, but I do not know the details of how this works in the case of psychopathy.

Because responsibility comes in degrees and every individual is different, it may seem that the theory of responsibility here offered is of little help in reaching concrete, generalisable answers to whether certain types of mental illness or disability should lead to criminal incapacity. This may be true, however, the theory does give some very helpful resources for considering specific cases. When we know what responsibility is, and are aware that it comes in degrees, we can consider specific cases on that basis. Given the exact information we have about a person in a specific case, do we have reason to think that the deliberation of this person could be influenced by being held morally responsible? If we take the Breivik case as an example, the approach in this article seems to be a more helpful and fruitful approach than the chaotic discussion of diagnoses that actually took place, where there was no principled discussion to explain why the specific diagnoses should imply capacity or incapacity of responsible behaviour.

5. Responses to Objections

In this section, I will respond to three main objections. Each could be the topic of a book on its own, but even if my answers are too brief to do the objections justice, a response is useful in order to clarify the theory further. The first objection is that a consequence-oriented approach to responsibility (like the one offered) fails to consider the relevance of justice to punishment. The second objection is that neuroscience shows that no people are responsible for their actions. The third objection is that philosophical arguments show that no people are responsible for their actions.

The first objection is as follows:[35] A consequentialist view of responsibility is prospective (forward-looking) while a desert-based view is retrospective (backwards-looking). The consequentialist view has the problem that it seems that quite contingent factors (like how likely it is that a person will change his behaviour) determines the responsibility that a person has. The desert-based view says, instead, that the blame or praise a person deserves is determined by looking back at what the person has done—on this view, what it means to say that a person deserves praise or blame is the same as saying that it would be just that the person got this amount of praise or blame.

I respond as follows: Above I distinguished between being responsible (having capacity for responsible behaviour) and being held responsible. When it comes to

[35] I thank David Vogt for this objection. He defends a retributive view in his doctoral dissertation: Vogt, *Crime, Punishment, and Understanding Justice through Injustice* (Doctoral dissertation, University of Bergen, 2018).

being held responsible, we should distinguish between being held responsible at all (being praised or blamed at all) and being blamed or praised to a certain degree or in a certain amount. Since I have said that being responsible comes in degrees, one could think that the amount of blame a person receives should correspond to the degree to which they are responsible. But that is not necessarily the case. Why not?

I am now discussing this question at a general level and not in individual cases. If people have a reduced capacity for responsible behaviour, should a consequentialist think that they should be blamed less or more for immoral behaviour? One could argue that they should be blamed less since they are less capable of behaving well and blaming them does not have the same effect as blaming someone more capable of changing behaviour. But one could also argue that they should be blamed more since it takes more to make them change their behaviour, or maybe because it is especially important to change their behaviour. For example, if their upbringing has made them more violent than normal, or they are born with a brain which makes them more aggressive or less capable of self-control, should they then be blamed more or less when doing something wrong?

The answer will here depend on the theory of ethics, theory of justice, and the understanding of humans that people have when they use an ethical standard for evaluating people's actions, as well as their general understanding of how to make the world a better place. These are ethical considerations that come in addition to the question about what responsibility is and what determines degrees of responsibility. While this is often considered to be part of what responsibility is, I argue that it is a separate question. Two persons may have the exact same consequentialist theory about what responsibility is and what determines the degree to which a person is responsible, and still, they can have a very different view on how strong blame or punishment a person deserves. That is because they have different views on how to weigh different consequences against each other and what are the best means to reach the desired goals.

It is a big question in meta-ethics whether moral norms have truth value, which cannot be discussed here. In any case—whether or not moral norms have truth value—individuals and societies must decide how strongly people should be blamed and punished, partly based on their capacity for responsible behaviour but partly also based on the specific goals that blame and punishment are intended to achieve (which may be more than what I have generally described here as an intention to cultivate better behaviour) and how efficient one thinks that different degrees of blame and punishment are as means to reach these goals.

After general answers have been given to these general questions, individual cases can be considered. Then we can ask whether the individual in question can be influenced, and to what degree it is self-chosen how much they can or cannot be influenced. Based on the general principles one adheres to when it comes to blame and punishment, one can deduce what to think about the individual case.

The second objection is from neuroscience. There are typically three kinds of findings which are used to argue that nobody is responsible for their own actions. The

first finding is Libet-style experiments showing that consciousness seems to enter the experienced deliberation process after the brain has already determined what a person will do.[36] More advanced experiments let researchers predict (better than chance) how people will act several seconds before they make their choice based on watching brain scans of the test persons.[37] The second finding is from confabulation-type and related kinds of experiments showing that our own conscious interpretations of our actions are often wrong. Confabulation means that persons are wrong about the real reasons for their actions. This has been demonstrated clearly in split-brain patients,[38] but also among people in general, typically in examples of choice blindness.[39] Other kinds of experiments show that non-conscious factors often influence our behaviour without us being aware of it.[40]The third finding is that reductionist theories of mind seem to explain all parts of human choices and actions.[41] Brain processes are physical processes guided by the laws of nature, and there is no need for concepts like persons with intentions choosing between alternatives and controlling the outcome.

I will not go deep into the neuroscience because the findings from neuroscience that are often considered to contradict free will in fact merely contradict specific theories of free will, like agent and non-causal libertarian theories. They lose their force if both the conscious and the non-conscious mind are understood as causal processes in any case. And they are even less relevant for theories of free will that emphasise free will as a result of decisions made over a long period of time, such as this one, since neuroscientific research on free will is usually made on spontaneous decisions. The reductive description that neuroscience offers of human choices are in full agreement with the descriptions that have been offered here.

The third kind of objection is philosophical. These will typically be objections which argue that humans do not have free will and thus cannot be held responsible for their actions either. I do not have space to discuss these objections here, but have done so at length in a book on free will.[42] Here I will only briefly explain how this model navigates different kinds of objections.

36 Libet, Freeman, and Sutherland, *The Volitional Brain: Towards a Neuroscience of Free Will* (Thorverton: Imprint Academic 1999).

37 Haynes et al., Reading Hidden Intentions in the Human Brain, *Current Biology* 17(4) (2007); Soon et al., Unconcious Determinants of Free Decisions in the Human Brain, *Nature Neuroscience* 11(5) (2008).

38 Gazzaniga (2005).

39 Johansson et al., Failure to Detect Mismatches between Intention and Outcome in a Simple Decision Task, *Science* 310(5745) (2005); Hall et al., How the Polls Can Be Both Spot on and Dead Wrong: Using Choice Blindness to Shift Political Attitudes and Voter Intentions, *PLOS ONE*, 8(4) (2013).

40 Schnall et al., Disgust as Embodied Moral Judgment, *Personality and Social Psychology Bulletin* 34(8) (2008).

41 Damasio (2010); Singer (2004).

42 Søvik (2016).

Some objections to free will (typically against compatibilists who think that free will is compatible with determinism) argue that we cannot be free if the world is determined, such as the manipulation argument and the zygote argument.[43] I think these are good arguments, but not against the theory presented here, since the theory presented in this article presupposes indeterminism at the macro level of human interaction. Some other objections to free will (typically against agent-causal or non-causal libertarians who reject that the mind is a causal process) is that free will presupposes appeals to mysterious agents or mysterious forms of causation, which does not fit into the ordinary scientific worldview. Nor do such agent-causal or non-causal theories explain how reasons make actions happen: in virtue of what does the agent control her actions?[44] The theory presented in this article does not appeal to a mysterious agent, but describes brain processes with the normal kind of causation, which fits well into an ordinary scientific worldview. Reasons cause actions, and this is the content of the concept of control.

The most relevant problems for the theory presented in this article, are the following three: 1) The problem of the disappearing agent, which says that when choices are reduced to desires, beliefs and bodily movement, the agent disappears; [45] 2) the regress problem, since it seems that if the mind is causal, we can follow causes backwards and backwards to before the agent can make an ultimate choice; [46] and 3) the luck argument, which says that (bad) luck is so pervasive in life that we cannot be said to have free will.[47] There is not space to discuss these objections here, but I mention them as relevant arguments and refer readers to how I have answered these objections in my book on free will, referenced above.

43 Pereboom (2014), pp. 76-79; Mele, *Free Will and Luck* (New York: Oxford University Press, 2006), pp. 188-89.

44 Pereboom (2014), pp. 65-69.

45 Helen Steward, *A Metaphysics for Freedom* (Oxford: Oxford University Press, 2012), 62.

46 Galen Strawson, 'The Impossibility of Moral Responsibility,' *Philosophical Studies* 75, no. 1/2 (1994): 5-7.

47 Neil Levy, *Hard Luck: How Luck Undermines Free Will and Moral Responsibility* (Oxford: Oxford University Press, 2011).